Between-speaker variability and temporal organization of the first formant

He, Lei ; Zhang, Yu ; Dellwo, Volker

DOI: https://doi.org/10.1121/1.5093450

Between-speaker variability and temporal organization of the first formant

Lei He, Yu Zhang, and Volker Dellwo

Citation: The Journal of the Acoustical Society of America **145**, EL209 (2019); doi: 10.1121/1.5093450
View online: https://doi.org/10.1121/1.5093450
View Table of Contents: https://asa.scitation.org/toc/jas/145/3
Published by the Acoustical Society of America

Between-speaker variability and temporal organization of the first formant

Lei He[a)]
Department of Linguistics, University of Tübingen, Wilhelmstrasse 19-23, DE-72074, Tübingen, Germany
lei.he@philosophie.uni-tuebingen.de

Yu Zhang and Volker Dellwo
Institute of Computational Linguistics, University of Zurich, Andreasstrasse 15, CH-8050, Zurich, Switzerland
yu.zhang@uzh.ch, volker.dellwo@uzh.ch

Abstract: First formant (*F*1) trajectories of vocalic intervals were divided into positive and negative dynamics. Positive *F*1 dynamics were defined as the speeds of *F*1 increases to reach the maxima, and negative *F*1 dynamics as the speeds of *F*1 decreases away from the maxima. Mean, standard deviation, and sequential variability were measured for both dynamics. Results showed that measures of negative *F*1 dynamics explained more between-speaker variability, which was highly congruent with a previous study using intensity dynamics [He and Dellwo (2017). J. Acoust. Soc. Am. **141**, EL488–EL494]. The results may be explained by speaker idiosyncratic articulation.

[AL]
Date Received: December 2, 2018 **Date Accepted:** February 19, 2019

1. Introduction

It has been repeatedly argued that speakers differ in articulatory movements.[1–6] Such articulatory idiosyncrasies leave traces in the acoustic signal, measurable in terms of speech rhythm,[1–3] formant trajectories,[4,5] and intensity dynamics.[6] This study explored how between-speaker differences are manifested in temporal organizations of the first formant (*F*1 hereinafter) with regard to *F*1 dynamics. We defined *F*1 dynamics as the speed of *F*1 increases from a minimum to its adjacent peak (positive *F*1 dynamics) and the speed of *F*1 decreases from a peak to its adjacent minimum (negative *F*1 dynamics). In other words, the average rate of *F*1 change per unit time was evaluated in both increasing and decreasing directions. *F*1 dynamics are a methodological fusion of McDougall[5] and He and Dellwo.[6]

Acoustic measures which can well capture between-speaker variability are essential in forensic phonetic practices. Both McDougall[5] and He and Dellwo[6] emphasized the importance of dynamic properties of the speech signal to investigating speaker-specific characteristics. Such idiosyncratic dynamic properties are typically associated with speaker-specific movement trajectories of articulators, which are a combined product of (i) idiosyncratic neurological substrates regulating the motor control of articulators, (ii) inborn anatomical peculiarities of speech organs constraining their biomechanics, and (iii) individual habits speakers acquired throughout their lifetime to operate articulators.[5–8] As a result, the acoustic characteristics in the speech signal, highly constrained by the articulatory dynamics, vary in speaker-dependent ways. These acoustic parameters include, *inter alia*, formant trajectories (modulated by, among other factors, the opening-closing gestures of the mouth [*F*1] and the fronting, backing, curling, and bunching gestures of the tongue [*F*2 and *F*3][9]), and intensity contours (co-varying with the mouth opening area as a function of time[10,11]).

In forensic phonetics, directly characterizing speaker-specific articulatory movements is almost impossible, as kinematic data of articulators are rare in trace materials. To crack this conundrum, forensic speech scientists focus on acoustic properties in the speech signal that are (although not entirely) modulated by the articulatory movements. For example, McDougall[5] approached the formant trajectories by using the least-squares polynomial approximations and found that the polynomial coefficients were useful for speaker discriminations. Alternatively, He and Dellwo[6]

[a)]Author to whom correspondence should be addressed. Also at: Institute of Computational Linguistics, University of Zurich, Andreasstrasse 15, CH-8050, Zurich, Switzerland.

measured the speeds of intensity increases and decreases between alternating peaks and troughs (i.e., intensity dynamics). Although the intensity fluctuations of the speech signal cannot be ascribed to a single factor, the opening–closing cycles of the mouth movements must play a non-trivial role. These cycles constantly change the geometry of the vocal tract, and accordingly its filter characteristics acting on the source signal, modifying its spectral properties and the intensity levels as a consequence. A high correlation between the signal intensity and the size of the mouth opening has already been demonstrated.[11] By calculating the speeds of intensity increases and decreases, it is possible to at least make some informed estimations of a speaker's idiosyncratic articulatory behavior. He and Dellwo[6] found that measures based on the speeds of intensity decreases (i.e., negative intensity dynamics) explained approximately 70% between-speaker variability, pointing to a possibility that the mouth-closing gestures may contain more speaker-specific information.

If the mouth closing gestures indeed encode more speaker-specific information, we should be able to obtain the same result using triangulated methods. In this paper, we analyzed the *F*1 trajectories, because, in addition to intensity, the mouth opening–closing movements also have an impact on the increases and decreases of *F*1.[12] Instead of fitting polynomials as in McDougall,[5] we calculated the *F*1 dynamics and tested whether measures of negative *F*1 dynamics also explained more between-speaker variability. Moreover, an advantage of using *F*1 over intensity is that *F*1 measures are less affected by varying distances to the microphone. This is particularly relevant in forensic scenarios, when voice experts typically have no information about the mouth-to-transducer distance.

2. Method

2.1 Corpus

The same TEVOID corpus[1,2] as used in He and Dellwo[6] was re-used in the present study. It contained 16 gender-balanced native speakers (mean age = 27, age standard deviation = 3.6, age range = 20–33; none reported speech and hearing disorders) of Zürich German (see Fleischer and Schmid[20] for a general phonetic description of the language). All speakers were recorded reading the same set of 256 sentences (13.2 syllables per sentence on average, standard deviation = 4.98 syllables per sentence) in a sound-attenuated booth through a mono channel (sampling rate = 44.1 kHz, quantization depth = 16 bits). The speakers practiced the sentences in advance to be able to read them fluently. They read the sentences in a way they considered "everyday reading." Boundaries of vocalic intervals were automatically demarcated based on manually tagged phonemes. A vocalic interval contained one monophthong or one diphthong acting as the syllable nucleus (altogether 3198 intervals were produced by each speaker). Sample recordings of the corpus was available in He and Dellwo[6] as supplementary materials.

2.2 Extracting the F1 curve and calculating F1 dynamics

For each vocalic interval, the *F*1 curve was extracted following the default Praat[13] routine, which included the following signal processing steps: (i) The signal was down-sampled to a Nyquist frequency of 5 kHz (for male voice) or 5.5 kHz (for female voice). (ii) The spectral slope above 50 Hz was pre-emphasized by an increase of 6 dB/octave. (iii) A Gaussian window (suppressing spectral skirts below −120 dB) was used to convolve the signal repeatedly (analysis window length = 25 msec, 3/4 between-window overlap). (iv) For each frame, the linear predicative coding (LPC) coefficients were computed using the Burg algorithm with ten poles; the *F*1 value (in Hertz) of this frame was pinpointed at the first peak in the LPC spectrum. Finally, the *F*1 curve of each vocalic interval for each speaker was linearly normalized within the range [0.01, 1] using the formula $F1'(m) = (1 - 0.01)/(\max - \min) \times [F1(m) - \min] + 0.01$, where $F1'(m)$ and $F1(m)$ refer to the normalized and original *F*1 curves, respectively, with the frame index m. Max and min refer to the old maximum and minimum values of $F1(m)$, and 1 and 0.01 are the new maximum and minimum values of $F1'(m)$. This way, the absolute range of vocal tract resonances was normalized to maintain the curvature of the original *F*1 trajectory only. The shape of the normalized *F*1 curve can now only be attributed to the idiosyncratic articulatory movements.

To measure the *F*1 dynamics, the peak *F*1 value ($F1_P$) and its associated time point (t_P) were obtained from the normalized *F*1 curve for each vocalic interval. Within each interval, the *F*1 minima to the left ($F1_{minL}$) and right ($F1_{minR}$) of the peak and their associated time points (t_{minL} and t_{minR}) were also obtained. Positive and negative *F*1 dynamics (abbreviated as *F*1[+] and *F*1[−], respectively) were defined

according to the formulas: $F1[+] = (F1_P - F1_{minL})/(t_P - t_{minL})$, and $F1[-] = |F1_{minR} - F1_P|/(t_{minR} - t_P)$. Geometrically, $F1[+]$ and $F1[-]$ can be visualized as the steepness of the secant lines $\overline{F1_{minL}F1_P}$ and $\overline{F1_PF1_{minR}}$ in Fig. 1.

To capture the distributions of both $F1[+]$ and $F1[-]$ in each sentence, the mean, standard deviation, and pairwise variability index [PVI; for an n-tuple $q_{n\geqq 3}$, its PVI is computed as $\sum_{i=1}^{n-1}|q_i - q_{n+1}|/(n-1)$[14]] were calculated. The PVI captures the averaged differences between adjacent acoustic magnitudes in a speech signal (e.g., duration,[2,14] intensity, and intensity dynamics,[3,6,15] or here $F1$ dynamics). It was demonstrated to be particularly suitable for summarizing the sequential variability in speech over the course of an entire utterance.[2,3,6,14,15] We notated these measures as mean_$F1[+]$, stdev_$F1[+]$, and pvi_$F1[+]$ for positive $F1$ dynamics, and mean_$F1[-]$, stdev_$F1[-]$, and pvi_$F1[-]$ for negative $F1$ dynamics. They represented different distributional aspects of $F1$ dynamics, namely, the central tendency, the overall dispersion, and sequential variability.

2.3 Data analyses

We used multinomial logistic regressions (MLRs) to test the significance of speaker and calculate the amount of between-speaker variability explained by the $F1$ dynamics measures. Measures of F1 dynamics were modeled as the numeric predictor variables, and speaker was modeled as the nominal response variable. For each MLR model, between-speaker variability explained by each measure was calculated in percentage points as $\chi^2/\sum\chi^2 \times 100$, where χ^2 refers to the likelihood ratio-χ^2 of a particular predictor, and $\sum\chi^2$ refers to the sum of χ^2's of all predictors in the model. Prior to model fitting, high-leverage values were scrutinized using Cook's distance (D). Multicollinearity among the $F1$ dynamics measures was examined using the scatterplot matrix. Data analyses were performed using JMP® 13.0 (SAS Institute Inc., Cary, North Carolina).

3. Results

No data points were found to be high-leverage ($D's < 1$). The scatterplot matrix (Fig. 2) revealed that multicollinearity existed among all positive dynamics measures (mean_$F1[+]$, stdev_$F1[+]$, and pvi_$F1[+]$) and all negative dynamics measures (mean_$F1[-]$, stdev_$F1[-]$, and pvi_$F1[-]$). As a result, we fitted three separate MLR models for each measure type (i.e., mean, stdev, or pvi); within each model, $F1[+]$- and $F1[-]$-based predictors were orthogonal. Table 1 shows the model fitting details and statistical results. For each of the three models, the measure of negative $F1$ dynamics explained ≈70% between-speaker variability, whereas the measure of positive $F1$ dynamics, ≈30% (see Fig. 3). In order to rule out the possible influence of phrase-final lengthening on the measures, the same procedure was repeated excluding the final vocalic interval in each sentence. Same results were yielded.

4. Discussion

This study investigated suprasegmental $F1$ dynamics in the speech signal. Results showed that measures of negative $F1$ dynamics explained about 70% between-speaker variability. Such results were highly congruent with our previous findings using

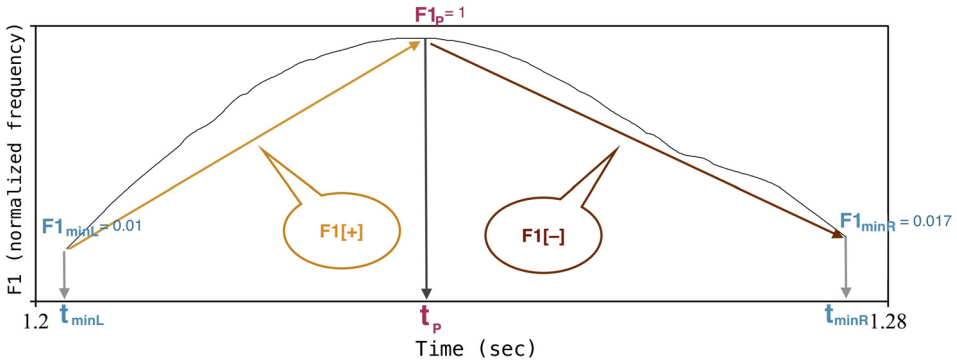

Fig. 1. (Color online) An illustration of calculating positive and negative $F1$ dynamics from a vocalic interval. The peak $F1$ value ($F1_P$), its flanking minimum values ($F1_{minL}$ and $F1_{minR}$), and their corresponding time points (t_P, t_{minL}, and t_{minR}) were extracted. A positive $F1$ dynamic ($F1[+]$) was calculated as the speed of $F1$ increased from an $F1$ trough to its adjacent peak, namely the steepness of the secant line $\overline{F1_{minL}F1_P}$. A negative $F1$ dynamics ($F1[-]$) was calculated as the speed of $F1$ decreased from an $F1$ peak to its adjacent trough, namely, the steepness of the secant line $\overline{F1_PF1_{minR}}$.

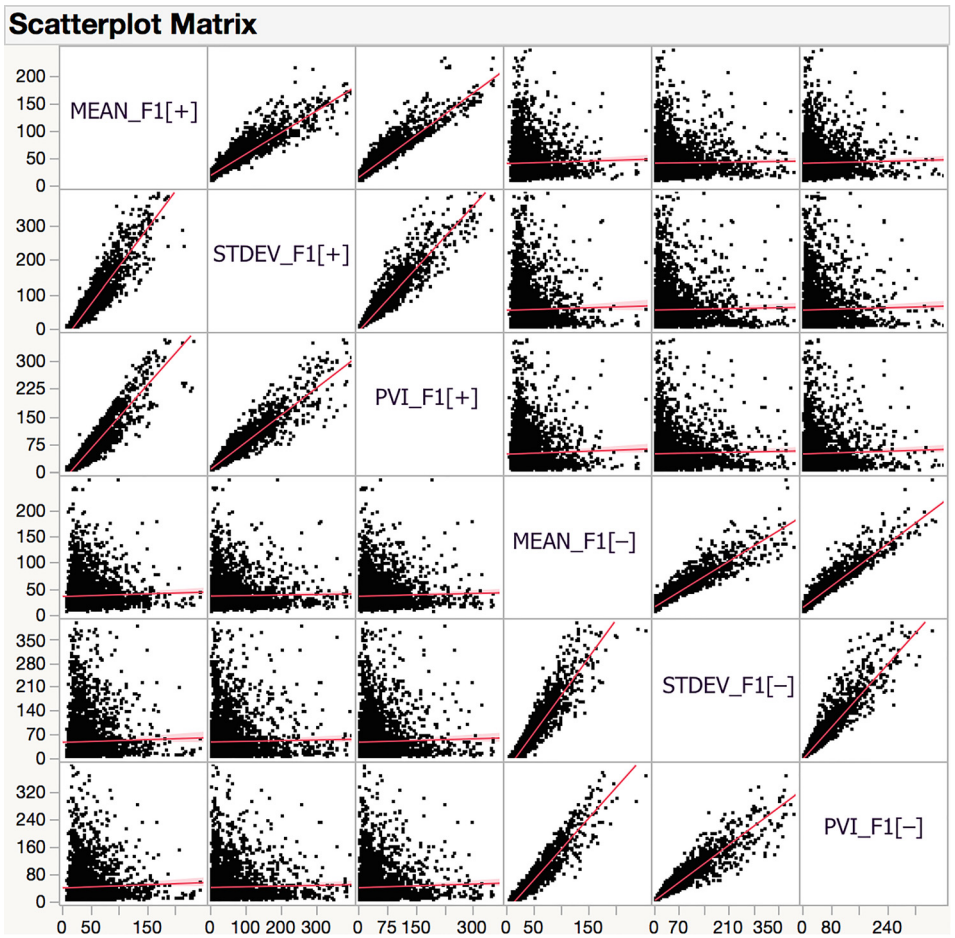

Fig. 2. (Color online) Scatterplot matrix with superimposed least-squares lines revealing that multicollinearity was present among measures of positive *F*1 dynamics (mean_*F*1[+], stdev_*F*1[+], and pvi_*F*1[+]) or measures of negative *F*1 dynamics (mean_*F*1[−], stdev_*F*1[−], and pvi_*F*1[−]). No multicollinearity existed between dynamics types ([+] or [−]).

Table 1. Results of MLRs.

	−logLik	$\chi^2_{(Df)}$[a]	P<	Variability explained[b]
(i) Model #1: Numeric predictors = mean_*F*1[+], mean_*F*1[−]; Response variable = speaker				
Full model	10 975.523			
mean_*F*1[+]-reduced model	11 093.044	$235.458_{(15)}$	0.0001	31.0%
mean_*F*1[−]-reduced model	11 236.984	$523.339_{(15)}$	0.0001	69.0%
		$\sum\chi^2 = 758.797$		$\sum\% = 100\%$
(ii) Model #2: Numeric predictors = stdev_*F*1[+], stdev_*F*1[−]; Response variable = speaker				
Full model	11 129.261			
stdev_*F*1[+]-reduced model	11 191.168	$123.814_{(15)}$	0.0001	27.3%
stdev_*F*1[−]-reduced model	11 293.937	$329.322_{(15)}$	0.0001	72.7%
		$\sum\chi^2 = 453.136$		$\sum\% = 100\%$
(iii) Model #3: Numeric predictors = pvi_*F*1[+], pvi_*F*1[−]; Response variable = speaker				
Full model	11 090.935			
pvi_*F*1[+]-reduced model	11 191.329	$159.464_{(15)}$	0.0001	30.0%
pvi_*F*1[−]-reduced model	11 276.402	$370.933_{(15)}$	0.0001	70.0%
		$\sum\chi^2 = 530.397$		$\sum\% = 100\%$

[a]The χ^2 value of each tested numeric predictor was calculated by taking twice the difference between the −logLik of the final model and the predictor-reduced model.

[b]The explained amount of between-speaker variability was calculated by taking the percentage of the χ^2 value of each predictor over the sum of the χ^2 values for both predictors ($\sum\chi^2$) in each model.

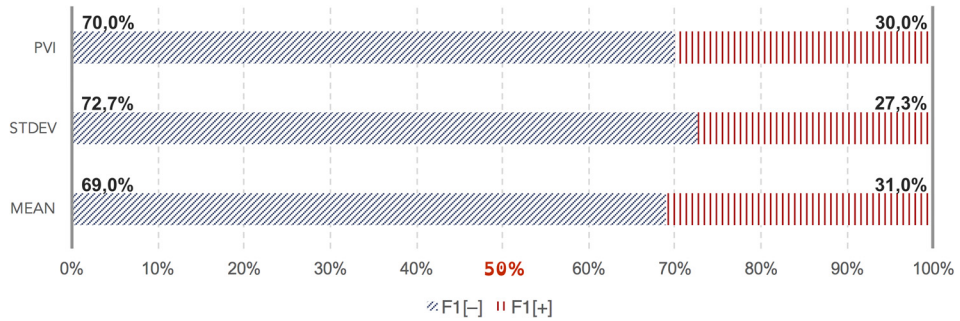

Fig. 3. (Color online) Bar chart illustrating the amount of between-speaker variability explained by measures of positive *F*1 dynamics (*F*1[+], vertical lines) and negative *F*1 dynamics (*F*1[−], diagonal lines) as per measure types (mean, stdev, and pvi).

intensity-based dynamics measures.[6] Why would measures of negative *F*1 dynamics contain more speaker-specific information? Similar to the intensity variations in the speech signal,[6] *F*1 covaries with the mouth opening and closing cycles.[12] These cycles are the articulatory basis of speech rhythm and play a crucial role in speech comprehension.[15,16] It is likely that in the mouth opening phase, the speaker may actively plan and control the articulatory movements to reach the peak in a syllable corresponding roughly to an *F*1 or intensity peak. In order to maximize mutual intelligibility, speakers should organize the articulatory movements in a temporally similar way while reaching the same peaks. Once the peaks have been reached, speakers may reduce the degrees of articulatory controls, producing movements controlled more by the inborn biomechanical properties of the osseous, muscular, and connective tissues involved in speech. The two properties of the motor plant,[17] controllable and intrinsic, might be differentially distributed in the mouth cycles. The controllable properties may be dominant in the opening phases, and the intrinsic properties may be dominant in the closing phases. A model-based study to reproduce the articulatory trajectories has suggested that the motor programs may be different in the opening and closing gestures.[18] Further research using articulatory measurements is demanded to testify our interpretation.

The findings of the present study have practical implications in forensic speech science. As our results indicated, speaker identity information is not uniformly distributed in the temporal organization of the *F*1 trajectories; the speeds of *F*1 decreases are more informative about a speaker's identity. This may mean that in forensic phonetic casework, experts should pay closer attention to the parts of the speech signal corresponding to the decreasing *F*1 regions, in case of interpreting formant dynamics. In addition, measures of the *F*1 dynamics may have potential advantages over other measures of formant dynamics, such as the polynomial curve fitting technique,[5] because the latter technique takes the whole trajectory into account, including parts with potentially less speaker identity information. On a practical level, our *F*1 dynamics measures are mathematically simpler than the polynomial coefficients, and are more readily explainable in terms of articulatory processes. This may be helpful for expert witnesses to present their analyses in front of the trier of fact, who may lack the respective mathematical background. Prior to applying our measures in casework, however, it is imperative to evaluate the speaker discrimination power of *F*1 dynamics, using the likelihood ratio framework[19] for instance. Regarding forensic applications, a caveat should be noted: our results were obtained from high-quality controlled recordings. It is imperative to investigate to what extent our method may be accurately applied to degraded quality recordings in forensic casework. For future research, it is also necessary to examine (i) the joint effects of other suprasegmental features on the *F*1 dynamics measures, including stress placement, speech rhythm, and intonation; (ii) the effect of spontaneous speech in different types of emotional valence on the *F*1 dynamics measures; and (iii) whether the same results can be replicated using other languages.

Acknowledgments

This work benefited from an early postdoc mobility grant (Grant No. P2ZHP1_178109) of the Swiss National Science Foundation to L.H. Portions of this work were presented at the 27th Annual Conference of the International Association for Forensic Phonetics and Acoustics and the International Conference on Laboratory Phonetics and Phonology.

References and links

[1]V. Dellwo, A. Leemann, and M.-J. Kolly, "Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors," J. Acoust. Soc. Am. **137**, 1513–1528 (2015).

[2]A. Leemann, M.-J. Kolly, and V. Dellwo, "Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison," Forensic Sci. Int. **238**, 59–67 (2014).

[3]L. He and V. Dellwo, "The role of syllable intensity in between-speaker rhythmic variability," Int. J. Speech, Lang. Law **23**, 243–273 (2016).

[4]K. McDougall, "Speaker-specific formant dynamics: An experiment on Australian English /ai/," Int. J. Speech, Lang. Law **11**, 103–130 (2004).

[5]K. McDougall, "Dynamic features of speech and the characterisation of speakers: Towards a new approach using formant frequencies," Int. J. Speech, Lang. Law **13**, 89–126 (2006).

[6]L. He and V. Dellwo, "Between-speaker variability in temporal organizations of intensity contours," J. Acoust. Soc. Am. **141**, EL488–EL494 (2017).

[7]V. Dellwo, P. French, and L. He, "Voice biometrics for speaker recognition applications," in *The Oxford Handbook of Voice Perception*, edited by S. Frühholz and P. Belin (Oxford University Press, Oxford, UK, 2018), pp. 777–795.

[8]P. Perrier and R. Winkler, "Biomechanics of the orofacial motor system: Influence of speaker-specific characteristics on speech production," in *Individual Differences in Speech Production and Perception*, edited by S. Fuchs, D. Pape, C. Petrone, and P. Perrier (Peter Lang, Frankfurt am Main, Germany, 2015), pp. 223–254.

[9]P. Ladefoged and K. Johnson, *A Course in Phonetics*, 6th ed. (Wadsworth, Boston, MA, 2011), xiv + 322 pp.

[10]Q. Summerfield, "Lipreading and audio-visual speech perception," Philos. Trans. R. Soc. London B **335**, 71–78 (1992).

[11]C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar, "The natural statistics of audiovisual speech," PLoS Comput. Biol. **5**, e1000436 (2009).

[12]D. Erickson, A. Suemitsu, Y. Shibuya, and M. Tiede, "Metrical structure and production of English rhythm," Phonetica **69**, 180–190 (2012).

[13]P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (version 6.0.32) [computer program]," http://www.fon.hum.uva.nl/praat/ (1992–2017) (Last viewed September 17, 2017).

[14]E. Grabe and E. L. Low, "Durational variability in speech and rhythm class hypothesis," in *Laboratory Phonology Seven*, edited by C. Gussenhoven and N. Warner (Mouton de Gruyter, Berlin, Germany, 2002), pp. 514–546.

[15]L. He, "Development of speech rhythm in first language: The role of syllable intensity variability," J. Acoust. Soc. Am. **143**, EL463–EL467 (2018).

[16]P. F. MacNeilage, "The frame/content theory of evolution of speech production," Behav. Brain Sci. **21**, 499–511 (1998).

[17]P. Perrier, "Gesture planning integrating knowledge of the motor plant's dynamics: A literature review for motor control and speech motor control," in *Speech Planning and Dynamics*, edited by S. Fuchs, M. Weirich, D. Pape, and P. Perrier (Peter Lang, Frankfurt am Main, Germany, 2012), pp. 191–238.

[18]P. Birkholz, B. J. Kröger, and C. Neuschaefer-Rube, "Model-based reproduction of articulatory trajectories for consonant-vowel sequences," IEEE Trans. Audio, Speech Lang. Process. **19**, 1422–1433 (2011).

[19]G. S. Morrison, E. Enzinger, and C. Zhang, "Forensic speech science," in *Expert Evidence*, edited by I. Freckelton and H. Selby (Thomson Reuters, Sydney, Australia, 2018), Chap. 99.

[20]J. Fleischer and S. Schmid, "Zurich German," J. Int. Phonetic Assoc. **36**, 243–253 (2006).